ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 5; Issue 2; March-April: 2017; Page No. 08-14

Automatic Text Summarization: Comparison of Various Techniques

S S R K Kiriti[1], Gogineni Siri[2], B Sai Priya[3], S Vijaya Durga Bhavani[4], Ch. Nanda Krishna[5]

[1,2,3,4] *IV/IV B Tech.* [5] *Professor, Department of Information Technology, VRSEC, Vijayawada, India*

sirigogineni4@gmail.com

Abstract

Transposing through multiple and large documents can be difficult and time consuming. So, this is when the concept of text summarization came into existence. Colossal documents are generally burdensome to summarize manually. Automatic text summarization is a process which enables us to reduce text with the help of a computer program in order to produce a summary which keeps possession of most important points of the original document. Over the past years, multitudinous techniques have come into existence which performs the automatic summarization such as term frequency, inverse document frequency, sentence position, sentence length, LSA and many more. In this paper, we compare these techniques and conclude which technique executes best.

Keywords: inverse document frequency, LSA, term frequency, text summarization;

Introduction

The process of enabling the user to understand a document by providing summary of the original document is called Text Summarization and if the summary is provided by means of a computer program then it is called Automatic Text Summarization. This summarization involves three steps. They are Pre-processing step, Processing step, Generation step. In the first step original text is obtained in a structured form. In the second step an algorithm is required to transfigure original text into its summary. In the last step the final summary is extracted from the original summary. There are two forms of summaries. One is Extractive Summary and the other is Abstractive Summary. Summarization methods are usually classified on the basis of levels of linguistic space in two detailed views. One is the Shallow approach and the other one is the deeper approach. The aim of the first approach is to reduce the space representation and it includes the removal of Stop Words, Case Folding and Stemming.

A. Extractive Summarization:

The computer itself selects some of the sentences or phrases or paragraphs based on some parameters like frequency, ranking etc. and give them as summary.

B. Abstractive Summarization:

Here we deal with the semantic analysis of the words. But developing such programs are hard because we need to implement Natural Language Processing techniques. The techniques which we are going to compare are:

a. Term Frequency(extractive)
b. Inverse Document Frequency(extractive)
c. Sentence Position(extractive)
d. Sentence Length(extractive)
e. Latent Semantic Analysis(abstractive)

After extracting the summary from all these techniques we will compare them with a judge summary which is generally an online summarizer. The technique which has the highest similarity percentage will be considered as the best among them. For identifying similarities we use cosine similarity method.

2. LITERATURE REVIEW

Over the past few years, there has been many methods used for the automatic summarization. Some of the methods are discussed.

It all started in 1958 by Luhn [1] who showed the significance of words based on frequency measures. The stop words are deleted and the rest words are given a hierarchy starting from root. The index describes the significance of each word. This calculation is done based on number of occurrences in the document and then ranked. Top sentences are

Corresponding author: S S R K Kiriti

selected based on this ranking and a summary is formed.

In 1961, G. J. Rath [2] determined relevant information from a set of documents using lexical indicators. Using this information the sentences are ranked and these ranked sentences are used for summarization.

In 1995, Julian Kupiec [3] determined different features like uppercase words, length, position of words by using naïve-bayes classifier by using a simple algebraic method. Single document summarization also became easy by the use of this method.

In 1997, Chin Yew Lin [4] used algebraic methods to determine the position of sentences. In the same year, Branimir Boguraev [5] considered an unstructured document; used saliency based content characterization to rank the important sentences. In 1999, Edward Hovy [6] used symbolic word knowledge with strong NLP techniques to show the concepts relevancies.

All the above findings are mainly about the single document summarization. Some of the multiple document summarization findings are mentioned below.
While coming to the multiple document summarizations McKeown and Raedev made major contribution in 1995 at Columbia university. They identified common themes using the extractive techniques and clustering.

In the same year 1995, Kathleen McKeon focussed on how the trends of events change with respect to time using the time-based technique and in 1997, Inderjeet Mani used graph based method to discover the nodes by applying a spreading activation technique.

In 2004, Jun'ichi Fukumoto [7] generated abstract by using Term Frequency (TF) and Inverse Document Frequency(IDF) for both single and multiple documents. In the same year, Rada Mihalcea added a vertex using graph based method for every sentence by creating links for similar sentences.

While coming to 2012, many were in a urge to bring out automatic text summarization using different methods. Vikrant Gupta used kernel to choose other sentences for summary by using statistical measures. Shanmugasundaram Hariharan extracted the sentences by using sentence co-relation method based on vote-casting, scores and positions to get extracts. On the other hand, Tiedan Zhu used the same sentence co-relation method and emphasized on logical-closeness rather than tropical-closeness.
This literature review mainly focussed on the work done by great personalities in the field of automatic text summar

3. PROPOSED SYSTEM

The proposed system for automatic text summarization and comparing various techniques is done by using the comparison between TF, IDF, Sentence Position, Sentence Length and Latent Semantic Analysis. All these techniques are applied on a single text document.

A. Simple Term Frequency (TF):

Term frequency, the word itself explains the number of occurrences of a term in a document. The high frequency words are considered here as required words in summary. The sentences that contain these high frequency words are ranked accordingly. But, there is a chance that due to the increase in the length of the document, the occurrence of unwanted top ranked sentences also exists thus making it difficult to operate on them. Also, high frequencies are difficult to be operated on. So, to overcome this, normalization is done which is nothing but mapping of higher values to lower values to the scale of 0 to 1. Formula for normalization is given as:

Normalized value of term $_i$ = frequency of term $_i$/ total number in the document But, there are still chances of considering unwanted top ranked sentences or not considering important low frequency words and corresponding sentences. So, to overcome this, a technique called Inverse Document Frequency (IDF) has to be performed.

B. Inverse Document Frequency (IDF):

This technique is mainly used for automatic text summarization which depends on the importance of a term in a document. There exists a formula for calculating IDF value:

IDF value of term i= log (total number of terms in a document/frequency of term i)

After getting both the Term Frequency (TF) and Inverse Document Frequency (IDF) values, they are multiplied to obtain a unique value (TF*IDF). It becomes difficult to perform operations if the obtained value is too high. So, normalization is

applied here also. Based on this normalized value, the sentences which contain the top preferred words are taken and the summary is formed accordingly. The main boon with this technique is that, along with the high frequency words, even the important low frequency words are also taken into consideration. But, there are also chances that the obtained words may not be related to the topic of the document. It is only useful as a lexical level feature. This is the main disadvantage with this technique. Thus, we go with the other technique, i.e. Sentence Position.

C. Sentence Position:

Sentence Position is one of the frequent techniques that are used in the text summarization. The general human's tendency is to place important topics of a document at certain positions in the document. The basic sentence position hypothesis is that the first few sentences in a document or a paragraph tells us the introduction about the document is the most important for summary and as the sentence gets further away from the beginning, it tells us about the description of the introduction and its importance also decreases (importance here means to place in summary). So, by using this technique, it is easy to carry out the summary of a document. This technique shows it's importance when we are dealing with journals, research papers, document that contain side-headings, and other topic related documents. But, this cannot be the best method as there may be cases where the first few sentences of a document may be less important than the latter. In such cases, this technique acts as a poor one.

D. Sentence Length:

The length of a sentence is generally expressed as number of words it contains. So, first we find out the sentences from the document and then we find the lengths of those sentences. Next, we calculate the maximum of the lengths. Next for every sentence we normalize the value by dividing its length with the length of the biggest sentence. Based on the maximum value of the length we fix a threshold value. The sentences containing length greater than that threshold value are not considered for summary. Generally not only longer sentences but also short sentences may not make sense some times. So, we also set a minimum value and consider only those sentences which have their length between minimum value and threshold value. The last technique which we aim to discuss and compare is the Latent Semantic Analysis (LSA).

E. Latent Semantic Analysis (LSA):

Generally, LSA is a natural language processing technique which analyses the relationships between a set of documents and the terms present in those documents. This is done by producing the concepts related to the terms and documents. This technique mainly assumes that words which have similar meaning occur in same piece of text. LSA is based on Singular Value Decomposition which is a mathematical matrix decomposition technique. It identifies both the relationships between the terms and sentences in an unstructured document and also determines the similarity of meanings between them. Firstly, the text document is taken as input and it acts as a matrix where each row represents the word and each column represents the column. The total value of the cell represents the importance of the word and in this paper, we use SVD technique to calculate this value. It forms an m*n matrix (M) and this matrix is formed as:

$M = A \sum V^T$

Where A is an m*n matrix which represents the rows as vectors of extracted values, $\sum$ represents a rectangular diagonal matrix with non-negative real numbers and V^T is an n*n real matrix which represents columns as vectors of extracted values. Lastly, after applying SVD to the input text document, the obtained result is used to select the sentences required to generate the summary.

These are the different techniques which are used in this paper for automatic text summarization.

After identifying summaries through all these methods we took an online summarizer as a judge summary and we found the similarities among the summaries. The similarity measure we took is Jacard Similarity.

F. Jaccard Similarity:

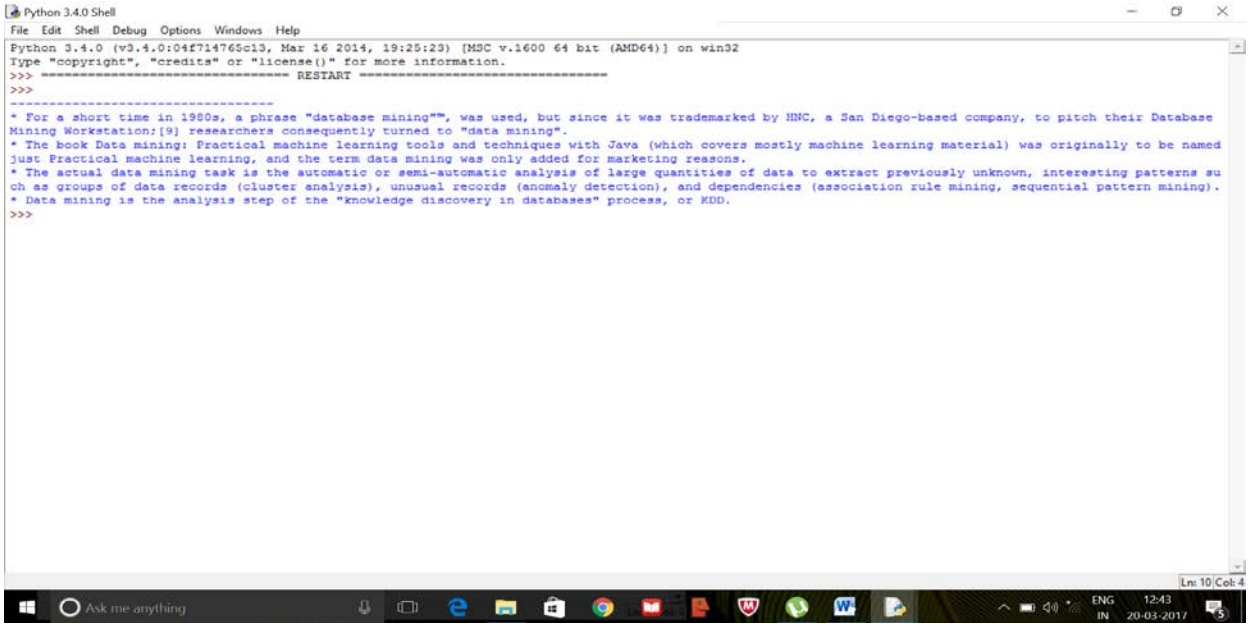
Consider two sets A = {0, 1, 2, 5, 6} and B = {0, 2, 3, 5, 7, 9}. The Jaccard similarity is defined as:

$JS(A, B) = |A \cap B| / |A \cup B| = |\{0, 2, 5\}| \; |\{0, 1, 2, 3, 5, 6, 7, 9\}| = 3/8 = 0.375$

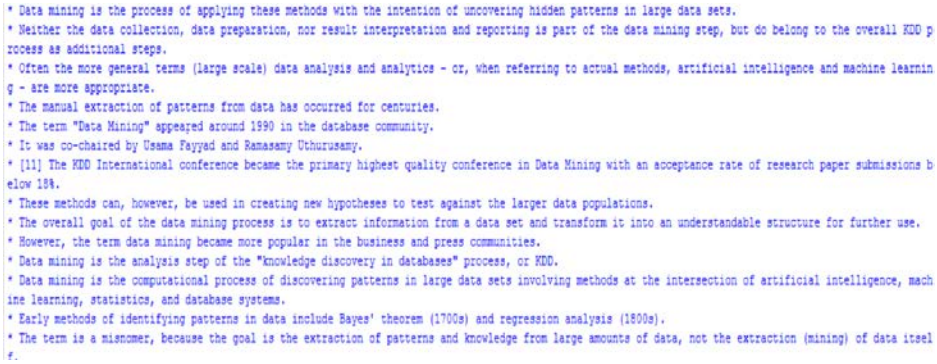
More notations, given a set A, the cardinality of A denoted |A| counts how many elements are in A. The intersection between two sets A and B is denoted $A \cap B$ and reveals all items which are in both sets. The union between two sets A and B is denoted $A \cup B$ and reveals all items which are in either set.

4. EXPERIMENTAL RESULTS AND OBSERVATIONS

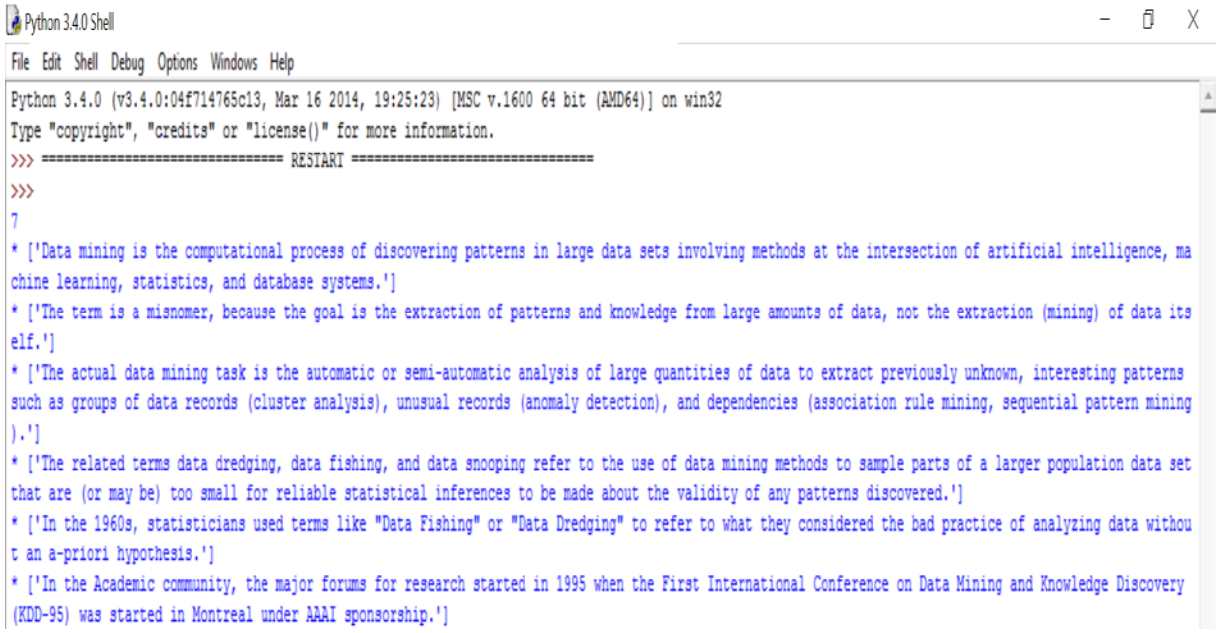
Term Frequency Summary



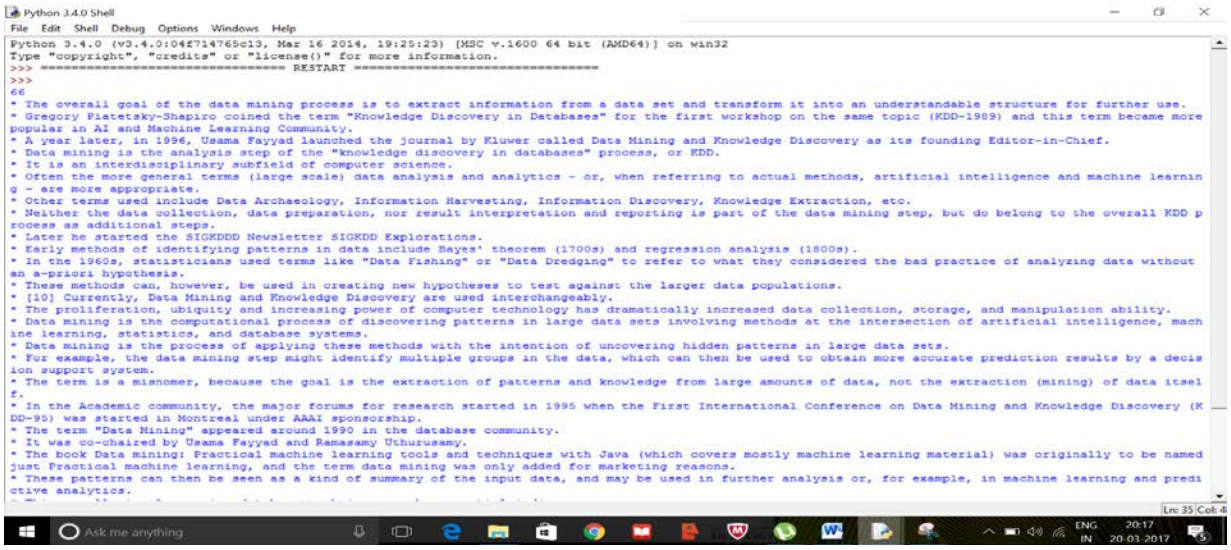
Inverse Document Frequency screenshot



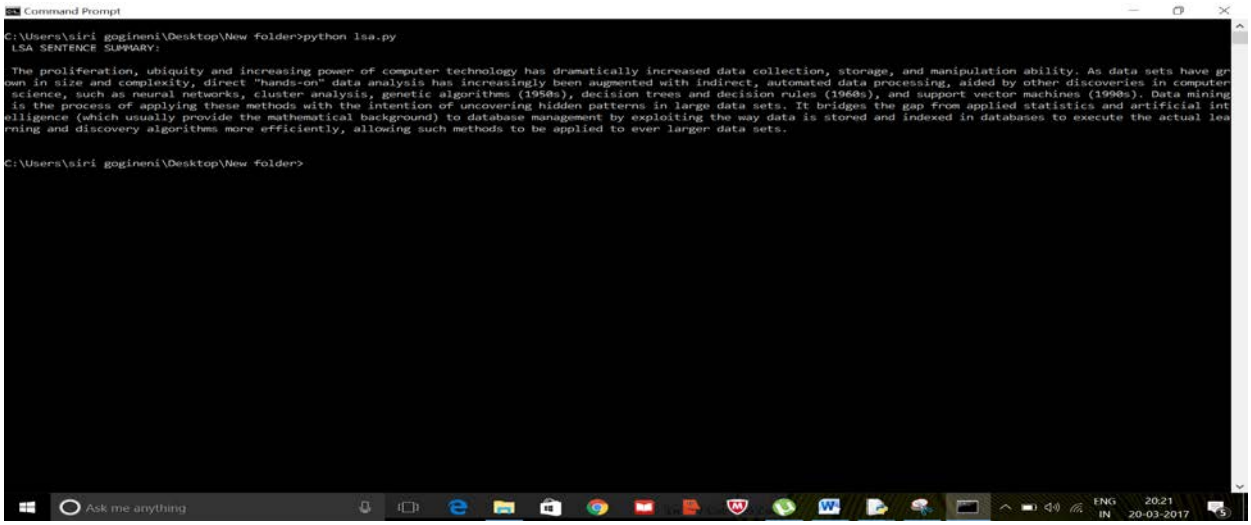
Sentence position screenshot



Sentence Length screenshot



LSA screenshot

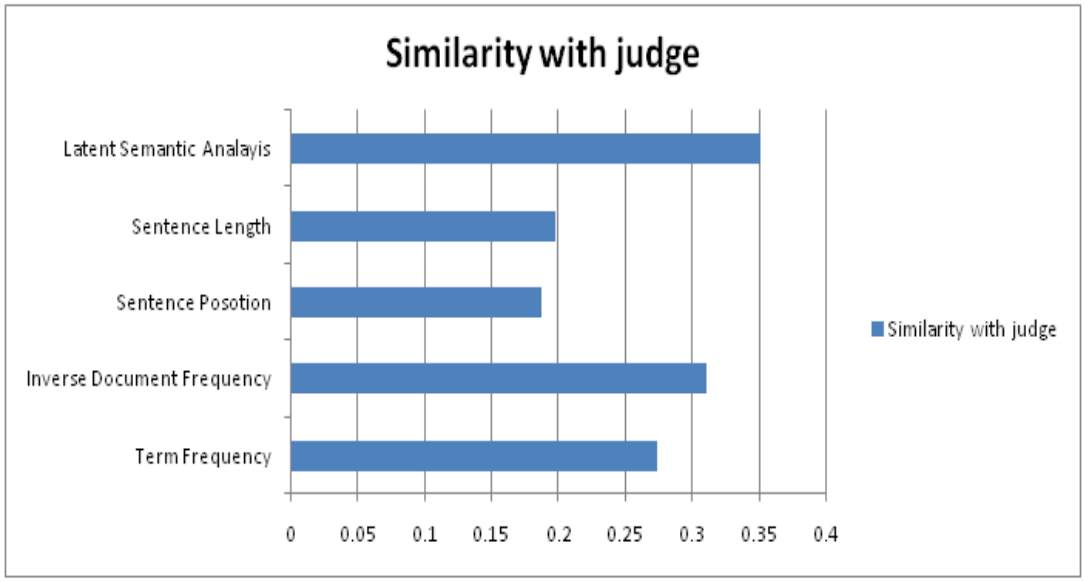


Jaccard Similarity screenshot

Similarity Table:

Technique	Similarity with Judge(0.0 to 1.0 scale)
Term Frequency	0.273291925
Inverse Document Frequency(tf-idf)	0.31
Sentence Position	0.187192118
Sentence Length	0.196891192
Latent Semantic Analysis	0.35

Similarity Graph:



5. CONCLUSION AND FUTURE WORK

Understanding a huge document without any abstract or summary is difficult and takes lot of time. This problem is solved with the automatic text summarization which identifies important sentences and words based on their frequency in a document and form a summary. Out of all the techniques used in this paper to perform automatic text summarization, Latent Semantic Analysis (LSA) is said to the best technique where semantically important sentences are identified. This paper mainly concentrates on single document. In future research, we plan to try all these techniques on multiple documents and compare which executes the best with optimum results.

REFERENCES

1. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159{165.
2. G. J. Rath, A. Resnick and T. R. Savage, “Comparisons of four types of lexical indicators of content,” Journal of the American Society for Information Science and Technology, vol. 12, no. 2, pp. 126- 130, April 1961.
3. Julian Kupiec, Jan Pedersen and Francine Chen, “A Trainable Document Summarizer,” Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 68- 73, 1995.
4. Chin-Yew Lin and Eduard Hovy, “Identifying Topics by Position,” In Proceedings of the Fifth conference on Applied natural language processing, San Francisco, pp. 283- 290, 1997
5. Branimir Boguraev and Christopher Kennedy, “Salience- based Content Characterization of Text Documents,” In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

6. Eduard Hovy and Chin-Yew Lin, Automated Text Summarization in SUMMARIST, In: Inderjeet Mani and Mark T. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, chapter 8, pp. 18- 24, 1999.

7. Jun'ichi Fukumoto, "Multi-Document Summarization Using Document Set Type Classification," Proceedings of NTCIR-4, Tokyo, pp. 412- 416, 2004.